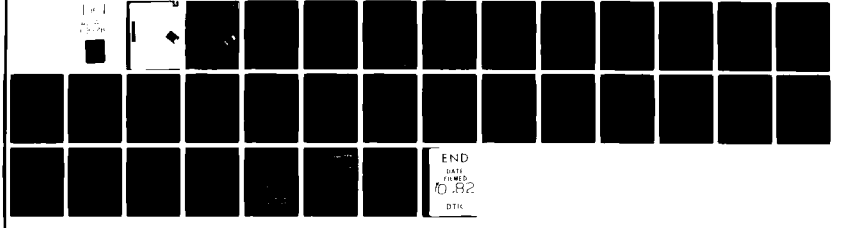
AD-A119 378 STANFORD UNIV CA DEPT OF STATISTICS F/G 12/1
A ROBUST ALTERNATIVE TO THE NORMAL DISTRIBUTION.(U)
JUL 82 D L MCLEISH N00014-76-C-0475
UNCLASSIFIED TR-321 NL



END
DATE FILMED
10-82
DTIC

# A ROBUST ALTERNATIVE TO THE NORMAL DISTRIBUTION

By

D. L. McLeish

TECHNICAL REPORT NO. 321

July 7, 1982

Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director





DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

# A Robust Alternative to the Normal Distribution

D. L. McLeish

Consider some of the arguments often advanced for the use of the normal distribution in a given situation:

(a) It is thought that the variable of interest can be represented as a sum of a large number of independent, small, and possibly identically distributed increments (i.e., the distribution is infinitely divisible).

(b) The distribution is symmetric and has all moments finite (it is argued that most real world measurements should have this property).

(c) The distribution is closed under convolutions; therefore useful for estimation and modelling.

(d) The normal distribution (or a reasonable facsimile) seems to fit many real world phenomena.

(e) The model can be extended to allow for dependent increments.

(f) The distribution is easy to handle. In the normal case, the maximum likelihood estimates are simple, although the distribution function requires numerical approximation. Generation of random normal variates is easy from a uniform generator.

On the other hand, many arguments have been made against the routine assumption of normality. Perhaps the most important of these is that "outliers" or "errors" seem to occur in otherwise normal samples and the normal estimates are highly non-robust to these. This observation has given rise to a considerable volume of literature in the robust theory of estimation. Many of the arguments and controversies surrounding, for example, the choice of the psi function for Huber's M-estimates are difficult for many to appreciate since they are presented either with

only heuristic justification or in reference to a seemingly artificial contamination model. The notion that there are "gross errors" in an experimenter's data is often one that is difficult to account for: furthermore, it is unnecessary to the justification of robust procedures as we shall see. Finally, although there is widespread agreement about the desirability of using robust methods, which such methods are most effective remains a controversial question (cf. Stigler (1977)). Some of the remaining reluctance to use robust methods may be related to Fisher's comment (quoted from Wilkinson (1979)):

"This example (relating to the Cauchy distribution) serves also to illustrate the practical difficulty which observers often find, that a few extreme observations appear to dominate the value of the mean. In these cases the rejection of extreme values is often advocated, and it may often happen that gross errors are thus rejected. As a statistical measure, however, the rejection of observations is too crude to be defended: and unless there are other reasons for rejection than mere divergence from the majority, it would be more philosophical to accept these extreme values, not as gross errors, but as indications that the distribution of errors is not normal. As we shall show, the only Pearsonian curve for which the mean is the best statistic for locating the curve, is the normal or Gaussian curve of errors. If the curve is not of this form the mean is not necessarily of any value whatever. The determination of the true curves for different types of work is therefore of great practical importance..."

The purpose of this paper is to discuss a family of distributions that seems more natural than the contamination models (for example, they

possess the properties (a) through (e) above), which approximate the normal distribution arbitrarily closely, and yet for which the maximum likelihood estimators are robust. Property (f) is not enjoyed by this family of distributions, since the simplest form for the density function in general is through a convergent power series. However, with the rapid decrease in the cost of high-speed computation, this defect is thought to be relatively unimportant. This distribution arises quite naturally in two different ways, and in a related paper is shown to provide a good fit to stock returns. Random generation of variates having this distribution requires only a normal and a gamma generator. The density function resembles that of the normal, having as support the whole real line, but the greater kurtosis makes this useful for modelling normal-like data in which there are apprarent "outliers".

The density function resembles that of the normal; it is symmetric, unimodal, and has support the whole real line. The kurtosis is greater than or equal to the normal, making this distribution useful in modelling phenomena with somewhat heavier tails than the normal. It has been used, for example, by Sichel (1973) to model the size distribution of diamonds.

This family connects two extreme members, the normal (for which the maximum likelihood estimates are not robust) and the Laplace (for which the maximum likelihood estimate of location is the median and is highly robust though not very efficient for normal-like distributions). There are other ways of connecting these two extremes. For example, the exponential power family (cf. Wilkinson (1979)) leads to $L_p$ estimates of location, $1 \leq p \leq 2$. For $p > 1$, these are not robust. Another family is proposed by O. Barndorff-Nielson (1977) to model the

distribution of sizes of sand particles. This distribution resembles ours in many ways, although it is not closed under convolutions. It could, like the present family, be used to derive robust estimates of location.

Some of the properties of the present family are also obtained by Teichroew (1957).

For convenience, distributions and their parameters will be as defined in Johnson and Kotz (1970).

The Density and Its Properties.

Let $G$ be a gamma distributed variate with parameters $(\alpha,2)$ and let $Z$ be a standard normal variate independent of $G$. Then the density function of $G^{\frac{1}{2}}Z$ is:

$$(1) \qquad g_\alpha(z) = \frac{1}{2^{\alpha+\frac{1}{2}}\sqrt{\pi}\Gamma(\alpha)} \int_0^\infty y^{\alpha-3/2}\, e^{-(z^2/2y)-y/2}\, dy \ .$$

This density is finite for all $z \neq 0$ if $\alpha > 0$ and finite for all $z$ if $\alpha > \frac{1}{2}$. For $\alpha > \frac{1}{2}$,

$$(2) \qquad g_\alpha(0) = \frac{\Gamma(\alpha - \frac{1}{2})}{2\sqrt{\pi}\Gamma(\alpha)} \ .$$

The modified Bessel function of the second kind is an even function that may be defined by:

$$K_\nu(z) = \tfrac{1}{2}\left(\frac{|z|}{2}\right)^{-\nu} \int_0^\infty t^{\nu-1}\, e^{-t-z^2/4t}\, dt \ .$$

If we now put $t = y/2$ in the definition of the density $g$, we obtain:

(3) $$g_\alpha(z) = \frac{1}{\sqrt{\pi}\Gamma(\alpha)}\left(\frac{|z|}{2}\right)^{\alpha-\frac{1}{2}} K_{\alpha-\frac{1}{2}}(z) .$$

This density was defined by Pearson, Jeffrey, and Elderton (1929) and investigated further by Pearson, Stouffer and David (1932). Here an asymptotic formula for large $\alpha$ is also given. The distribution is applied to testing for differences between chi-squared values in contingency tabled data, and tables of the distribution function for small values of $\alpha$ are provided. This density may be used to describe the sample covariance between independent, identically distributed normal samples.

It is not difficult to show that this family of densities satisfies, for positive a , a homogeneous differential equation of the form

$$g''_\alpha(z) - \frac{2(\alpha - 1)}{z} g'_\alpha(z) - g_\alpha(z) = 0 .$$

This obtains from the modified Bessel equation satisfied by $K_\nu(z)$ ,

$$z^2 K''_\nu + z K'_\nu - (z^2 + \nu^2) K_\nu = 0 .$$

$g_\alpha(z)$ also satisfies the difference equation:

$$\frac{z^2}{2(\alpha - 1)} g_{\alpha-1}(z) - 2\alpha g_{\alpha+1}(z) + (2\alpha - 1) g_\alpha(z) = 0$$

which follows from the equation:

$$z\,K_{\nu-1}(z) - z\,K_{\nu+1}(z) = -2\nu K_{\nu}(z) \ .$$

A more general family:

$$g_{\alpha,a}(x) = (1 - a^2)^{\alpha}\, e^{ax}\, g_z(x)$$

may also be defined. This is closely related to the Bessel function distribution (cf. McKay (1932) and Laha (1954)). In this model, however, it is fairly difficult to disentangle the parameters $\alpha,a$ from the location-scale parameters. We will therefore concentrate on (3) although many of its properties carry over to its generalization (cf. Press (1967)).

We will adopt (3) as the definition of the density since this will allow the function to be defined even for negative values of $\alpha$ (although in this case it is not a density function).

We now introduce location and scale parameters to obtain the more general family of densities:

$$(4) \qquad f(x;\mu,\theta,\alpha) = \frac{1}{\theta}\, g_{\alpha}\left(\frac{x-\mu}{\theta}\right) \ .$$

We express this general family as $Be(\mu,\theta,\alpha)$. This is the density function of a constant $\mu$ plus the product of a standard normal variate with one having the distribution of the square root of a gamma $(\alpha,2\theta^2)$ variate. The moment generating function of the density is:

$$(5) \qquad m(t;\mu,\theta,\alpha) = e^{\mu t}(1 - \theta^2 t^2)^{-\alpha} \ .$$

Since (5) factors into $e^{\mu t}(1-\theta t)^{-\alpha}(1+\theta t)^{-\alpha}$, it is easily seen that this is also the density of $\mu + G_1 - G_2$ where $G_1$ and $G_2$ are independent gamma $(\alpha,\theta)$ variates. Moreover, since $m(t;\mu_1,\theta,\alpha_1)\, m(t;\mu_2,\theta,\alpha_2) = m(t;\mu_3,\theta,\alpha_3)$ where $\mu_3 = \mu_1 + \mu_2$ and $\alpha_3 = \alpha_1 + \alpha_2$, this distribution is closed under convolutions (for fixed scale parameters) and is therefore infinitely divisible. It therefore has a representation such as that in property (a) above. The central moments of the distribution are:

$$E|X-\mu|^{2p-1} = (2\theta)^{2p-1}\,\frac{\Gamma(p)\,\Gamma(\alpha+p-\tfrac{1}{2})}{\pi^{\frac{1}{2}}\Gamma(\alpha)}$$

and in particular, when $2p - 1 = 1, 2, 4$ respectively, we obtain

$$\frac{2\theta\Gamma(\alpha+\tfrac{1}{2})}{\sqrt{\pi}\Gamma(\alpha)}\ ,\quad 2\theta^2\alpha\ ,\quad \text{and}\quad 12\theta^4\alpha(\alpha+1)\ .$$

Two important cases of this family of densities deserve mention. The first is the case $\alpha = 1$, when

$$g_1(z) = \tfrac{1}{2}e^{-|z|}\ . \tag{6}$$

This is the Laplace distribution, for which the maximum likelihood estimates of location and scale are sample median and average absolute deviation. The second case occurs when, for $\sigma^2$ a positive constant, $\theta$ approaches 0 and $\alpha$ approaches infinity. Specifically, as $\theta \to 0$, $f(x;\mu,\theta,\sigma^2/2\theta^2)$ approaches the normal $(\mu,\sigma^2)$ density.

When $\alpha$ is an integer, say $\alpha = n + 1$, $n > 0$, we have the following representation of $g$.

$$g_\alpha(z) = \sum_{k=0}^{n} \frac{|z|^k}{k!} e^{-|z|} \binom{2n-k}{n} \frac{1}{2^{2n-k+1}} . \tag{7}$$

This is clearly a convolution between Poisson probabilities and those as a negative binomial type, i.e.,

$$p_j = \binom{n+j}{n} \frac{1}{2^{n+j+1}} , \quad j = 0, 1, 2, 3, \ldots . \tag{8}$$

Indeed, if $X$ has a Poisson distribution with parameter $|z|$ and $Y$ has probabilities specified by (8), then

$$g_\alpha(z) = P(X+Y=n) .$$

The first few densities of this type are easily written out:

$$g_1(z) = \tfrac{1}{2}e^{-|z|} \qquad g_3(z) = (z^2 + 3|z| + 3)e^{-|z|}/16$$

$$g_2(z) = \tfrac{1}{4}(1 + |z|)e^{-|z|} \qquad g_4(z) = \frac{e^{-|z|}}{32}\left\{5 + 5|z| + 2z^2 + \frac{|z|^3}{3}\right\} .$$

See Figure 1 for graphs of the densities. The distribution function may be defined by symmetry and the relation:

$$P(|X| \le x) = 2 \sum_{k=0}^{n} p_{n-k} \frac{\gamma(k+1,x)}{k!} \tag{9}$$

where $\gamma(a,x)$ is the incomplete gamma function $= \int_0^x t^{a-1} e^{-t}\, dt$.

For the purpose of estimating location or scale, it is often convenient to generate the score function directly. Setting

$$\eta_\alpha(x) = \frac{-g'_\alpha(x)}{g_\alpha(x)} ,$$

and using the difference equation satisfied by $g_\alpha$ ,

$$\eta_{\alpha+1}(x) = \frac{x}{x\,\eta_\alpha(x) + (2\alpha-1)}$$

and for integral $\alpha$ , we have a continued fraction expansion,

$$\eta_{\alpha+1}(x) = \left[\frac{2\alpha-1}{x} + \ldots + \left[\frac{3}{x} + \left[\frac{1}{x} + \operatorname{sgn} x\right]^{-1}\right]^{-1}\right]^{-1} .$$

Then, given $g_\alpha$ for small $\alpha$ , we may generate subsequent densities through the expression

$$g_{\alpha+m}(x) = \left(\frac{x}{2}\right)^m g_\alpha(x)\,\frac{\Gamma(\alpha)}{\Gamma(\alpha+m)} \prod_{k=1}^{m} \eta^{-1}_{\alpha+k}(x) . \tag{10}$$

Here, $\Gamma(k)$ is the complete gamma function, $\gamma(k,\infty)$. This approach is often preferable to the use of the series expansion valid when $\nu = \alpha - \frac{1}{2}$ is not an integer:

$$g_\alpha(x) = \frac{\sqrt{\pi}}{2\Gamma(\alpha)\sin(\nu\pi)} \sum_{k=0}^{\infty} \frac{\left(\frac{z}{2}\right)^{2k}}{k!} \left[\frac{1}{\Gamma(-\nu+k+1)} - \frac{\left(\frac{|z|}{2}\right)^{2\nu}}{\Gamma(\nu+k+1)}\right] .$$

An alternative expansion to (7), useful for large $|z|$ , and $\alpha$ integral is the following:

$$g_\alpha(z) = \frac{|z|^{\alpha-1}}{\Gamma(\alpha)\, 2^\alpha} e^{-|z|}\left\{1 + \frac{\mu-1}{8|z|} + \frac{(\mu-1)(\mu-9)}{2!(8z)^2} + \frac{(\mu-1)(\mu-9)(\mu-25)}{3!(8|z|)^3} + \ldots\right\}$$

where $\mu = 4(\alpha - \frac{1}{2})^2$.

An expression for the $\frac{\partial}{\partial\alpha} g_\alpha(z)$ useful for maximum likelihood estimation, is obtained by differentiating (10):

$$\frac{\partial}{\partial\alpha} g_\alpha(z) = -\psi(\alpha)\, g_\alpha(z) - \pi\cot(\nu\pi)\, g_\alpha(z)$$

$$+ \frac{\sqrt{\pi}}{2\Gamma(\alpha)\sin(\nu\pi)} \sum_{k=0}^{\infty} \frac{\left(\frac{z}{2}\right)^{2k}}{k!}\left[\frac{\psi(-\nu+k+1)}{\Gamma(-\nu+k+1)}\right. \tag{11}$$

$$\left. + \frac{\left(\frac{z^2}{4}\right)^{\nu}}{\Gamma(\nu+k+1)}\left\{\psi(\nu+k+1) - \ln\frac{z^2}{4}\right\}\right]$$

where $\psi(\alpha)$ is the digamma function $\frac{d}{d\alpha}\ln\Gamma(\alpha)$.

Estimation of Parameters.

Before we present the maximum likelihood estimators of the parameters, we require an elementary property of the score function that is based on the similar property for the Bessel function (cf. Abramowitz and Stegun (1964)): $zK_\nu'(z) + \nu K_\nu(z) = -zK_{\nu-1}(z)$ for all $\nu$. Therefore,

$$\frac{g_\alpha'(z)}{g_\alpha(z)} = -\frac{K_{\nu-1}(z)}{K_\nu(z)} = -\frac{z}{2(\alpha-1)}\frac{g_{\alpha-1}(z)}{g_\alpha(z)} \tag{12}$$

for $\alpha > 1$ and $z \neq 0$.

We now consider the problem of estimating the parameters. We start with the estimation of $\alpha$. To begin with, as $z \to \infty$,

$$g_\alpha(z) \sim \frac{1}{\Gamma(\alpha)2^\alpha} z^{\alpha-1} e^{-z} ,$$

and

$$\frac{\frac{\partial g_\alpha(z)}{\partial \alpha}}{g_\alpha(z)} \sim -\psi(\alpha) + \ell n \frac{|z|}{2} .$$

Therefore, maximum likelihood estimation of $\alpha$ for large $z$ is nearly achieved by setting the sample mean of the variables $\ell n\, z_i^2$, $i = 1, 2, \ldots, n$ equal to their expected value and solving for $\alpha$. Moreover, since the shape parameter $\alpha$ is primarily evident for $\alpha > 3/2$ by the weight in the tails, and since true maximum likelihood estimation of $\alpha$ is fairly cumbersome computationally, this is one approach we used. When true maximum likelihood estimation was attempted on samples, the iterates often seemed to fail to converge. Furthermore, we are primarily interested here in efficient estimation of $\mu$, and as we shall see, this depends very little on getting an accurate estimate of $\alpha$.

One possibility for the estimator of parameters is the simple moment method. For example, one approach tried was to estimate $\alpha$ from the relation:

$$E \log(X_1-\mu)^2 = -\gamma + \log \theta^2 + \psi(\alpha)$$

where $\gamma$ is Euler's constant .57721... and $\psi$ is the digamma function.

This equation was solved recursively for $\alpha$ after initial estimation of $\mu$ and $\theta$.

Another approach that is feasible though reasonably expensive is full maximum likelihood estimation of all three parameters. While this may be an asymptotically desirable procedure, or useful if we are interested in accurate estimates of $\alpha$, it is not necessarily the best method when attention is focused on estimating $\mu$ for small sample size. In fact, some evidence was obtained that full MLE had lower efficiency for the estimator of $\mu$ ($n = 20$, underlying distribution = normal) than the simple minded scheme used in the simulation.

The rather difficult question of which scheme for estimating $\alpha$ leads to highest efficiency for estimation location will not be dealt with definitively here. However, after estimating $\alpha$, we may estimate $\mu$ and $\theta$ as follows:

The maximum likelihood equation for $\mu$ is:

$$\sum_{i=1}^{N} \frac{g'_\alpha(z_i)}{g_\alpha(z_i)} = 0 \quad \text{with} \quad z_i = \frac{X_i - \mu}{\theta}$$

or, if $\alpha > 1$ and no $z_i = 0$,

$$\mu = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i} \tag{14}$$

where $w_i = g_{\alpha-1}(z_i)/g_\alpha(z_i)$. Similarly, the maximum likelihood equation for $\theta$ is:

$$\sum_{i=1}^{N} z_i g'_\alpha(z_i)/g_\alpha(z_i) + N = 0$$

or, if $\alpha > 1$ ,

$$K = (\alpha-1)\theta^2 = \frac{1}{2N}\sum_{i=1}^{N} w_i(X_i-\mu)^2 \quad . \tag{15}$$

Now (14) and (15) are iterated for fixed $\alpha$ until convergence occurs.

When $\alpha$ is known, the Fisher information matrix is given by:

$$I = \frac{1}{4\theta^2(\alpha-1)^2}\begin{pmatrix} \int_{-\infty}^{\infty} z^2 g_{\alpha-1}^2(z)/g_\alpha(z)\, dz & 0 \\ 0 & 4(\alpha-1)^2 + \int_{-\infty}^{\infty} z^4 g_{\alpha-1}^2(z)/g_\alpha(z)\, dz \end{pmatrix}.$$

Some features of these equations are interesting in the light of the theory of robust estimation. For the purpose of this discussion, let us assume $\alpha > 1.5$. Note by (2) that when $\alpha > .5$, $g_\alpha(0) = \Gamma(\alpha-\frac{1}{2})/2\sqrt{\pi}\Gamma(\alpha)$. Therefore, corresponding to $z = 0$ , we assign weight $w_i = (\alpha-1)/(\alpha-3/2)$.

Similarly, as $z \to \infty$, $g_\alpha(z) \sim \frac{1}{\Gamma(\alpha)2^\alpha} z^{\alpha-1} e^{-z}$ and so corresponding to large $z_i$ ,

$$w_i \sim 2(\alpha-1)/z_i \quad .$$

This produces the effect that the influence curve is approximately linear in the central region, but bounded for all $z$. This is true for any finite $\alpha$ although the influence curve for the normal is unbounded and densities in this family can approximate arbitrarily closely the normal density by taking $\alpha$ sufficiently large.

Thus maximum likelihood estimation within this family seems to be highly robust, but of course, robustness may be achieved at considerable loss of efficiency. Robustness against obtaining a false value of $\alpha$ is particularly important here, since although the mean $\mu$ and the variance $2\theta^2\alpha$ are relatively easy to estimate, the estimate of $\alpha$ itself is subject to considerable error. We considered only the loss in efficiency in the estimation of the location parameter $\mu$ when $\alpha$ is misspecified. To take an extreme case, let us suppose that the observations arise from a normal distribution ($\alpha=\infty$) with mean $\mu$ and known variance $\sigma^2 = 10$. Let us suppose, however, we believe $\alpha = 5$ (so the value of $\theta^2$ is $.1\sigma^2$). Then the asymptotic relative efficiency of the ML estimator of $\mu$ obtained from the solution of (14) is

$$\frac{\sigma^2[E\ \phi'(Z)]^2}{E\ \phi^2(Z)}$$

where $\phi(x) = \dfrac{g_\alpha'(x)}{g_\alpha(x)}$ and $Z$ is $N(0,\sigma^2)$. Figure 2 shows the asymptotic efficiency of the estimator for a normal sample when various other values of $\alpha$ are assumed. Note that there is high efficiency for any value of $\alpha$ above about 2 and the worst value ($\alpha = 1$ for the sample median) is .637. The efficiency seems to be much flatter as a function of sample size than for many other robust procedures. Holland and Welsch (1977) show that when the scale parameter is estimated, the small sample efficiency for many of the robust procedures seems to be substantially below its asymptotic value. For comparison, we used $\alpha = 3.46$ (asymptotic efficiency = 95%) and Monte Carlo methods on normal samples of size 10 to obtain an estimate of the efficiency for $n = 10$ of .94.

On the other hand, if the observations come from a distribution of this form with $1 \leq \alpha < \infty$, we may choose to use $\alpha = 1.7$ in constructing the estimates. This provides asymptotic efficiency for location at least 84.7% for all the members of this family.

We now discuss briefly the asymptotic efficiency of these estimates for the Cauchy distribution, a large tailed distribution which is clearly not a member of this family. Consider an M-estimate in general, obtained as the solution of:

$$\sum_i \eta(\lambda(x_i-\mu)) = 0 \ . \tag{16}$$

The minimum asymptotic variance over all choices of the scale parameter $\lambda$ is:

$$\min_{\lambda} \frac{\int_{-\infty}^{\infty} \eta^2(\lambda x)\ F(dx)}{\lambda^2\left\{\int_{-\infty}^{\infty} \eta'(\lambda x)\ F(dx)\right\}^2} \ . \tag{17}$$

Consider $F$ to be the Cauchy distribution function (in standard form) and $\eta(x) = x(1-x^2)^2$ for $|x| < 1$, 0 otherwise. This is Tukey's bi-square and results in maximum asymptotic efficiency obtained from (17) of about 90%. This efficiency is quite sensitive to the value of $\lambda$. As $\lambda$ increases beyond its optimal value, the efficiency decreases to 0 because of the "redescending" nature of the function $\eta$. Therefore, a reasonably accurate estimate of the scale factor $\lambda$ is critical for efficient estimates of location. Replacing $\eta(x)$ by $g_\alpha'(x)/g_\alpha(x)$ results in asymptotic efficiency of around .86 for $\alpha = 2$ and 3, for example. In this case, however, the asymptotic efficiency is flat for

$\lambda$ in a broad neighbourhood of the optimum and as $\lambda \to \infty$, it approaches $8/\pi^2$. Here, the choice of the optimal $\lambda$ is not nearly as critical and results in the better performance of the estimator for $n = 20$ in the simulations of the next section. Bell (1980) explores problems in adaptive estimation of the optimal $\lambda$ for the bisquare.

It may be desired to build in more adaptivity for large tailed distributions into this family. In fact, the family could be broadened to include distributions whose tails decay at a slower than exponential rate (and therefore lead to redescending score functions) by replacing the gamma distribution which multiplies the normal by a distribution such as Fisher's F. Of course, the resulting distribution will no longer be closed under convolutions, one of the most attractive features of the present family.

How much "robustness" is purchased with the loss in asymptotic efficiency evident in Figure 2? Since the influence curve is proportional to the score function $\frac{g_\alpha'(x)}{g_\alpha(x)}$ which is bounded, the estimates are robust. This function is also smooth for $\alpha$ adequately greater than 1. Treatment of "outliers" is evident from the score function graphed in Figure 3 for various values of $\alpha$. These functions are all asymptotic to 1 as $x \to \infty$. Consider, for example the case $\alpha = 2$. It is seen that the value of the score function at $x = 1$ (roughly .7 standard deviations from the mean) is around .5. In other words, in iterating (14), one observation at $\infty$ is only the equivalent of about 2 observations at $x = 1$. The degree of robustness for the other values of $\alpha$ may similarly be compared.

Small Sample Behavior.

Andrews et al. (1972) indicate that adaptive estimates do not normally outperform non-adaptive ones except for moderate and large sample sizes. It is therefore not expected that these estimates can consistently outperform others over a broad range of distributions. On the other hand, the attractive features of this family of distributions leads one to hope that it does not suffer from the lack of robustness of normal estimators, or from too great a lack of efficiency by comparison with others. There is little doubt that these estimators will be competitive with the trimmed mean or Huber's proposal 2; the influence functions can be made similar and the distributions considered here, like Huber's least favourable distribution, all have exponentially decreasing tails. It was therefore decided to make the comparison on possibly unfavourable ground; with an M-estimate of redescending type such as Tukey's bisquare, and including wide tailed distributions such as the Cauchy and the slash. We also chose a relatively small sample size, $n = 20$ for the comparison. Adaptive estimates are usually expected to improve their performance for increasing $n$.

Exact maximum likelihood estimation of all three parameters is computationally feasible and was performed on several samples of $n = 20$. Some problems are apparent because of the very flat nature of the likelihood as a function of $\alpha$. For example, a significant proportion of normal samples ($\alpha = \infty$) leads to MLE of $\alpha$ constrained to the interval $[1,\infty]$ of 1 and the resulting estimate of location the sample median. This seems to result in loss of efficiency for the normal. Due to the cost of full MLE, it was impossible to conduct a simulation

study of the performance of the estimators and so the following crude substitute was used. It seemed to provide comparable efficiency for the estimation of location to a scheme based on likelihoods (but not, of course, for estimation of $\alpha$).

Define $m = \text{med}(X_i)$ and

$$k = \frac{n \sum_{i=1}^{n} (X_i - m)^4}{\left[\sum_{i=1}^{n} (X_i - m)^2\right]^2} .$$

The estimator of $\alpha$ was

$$\hat{\alpha} = \begin{cases} 1 , & k \geq 6 \\ 2 , & 4.2 \leq k < 6 \\ 3 , & k < 4.2 \end{cases}$$

In the case $\hat{\alpha} = 1$, the scale parameter does not affect the location estimator which is simply the sample median. In case $\hat{\alpha} = 2$ or $3$, the scale parameter was estimated crudely by matching the median absolute deviation of the sample with that of the assumed distribution. Setting $\text{MAD} = \text{med}|X_i - \text{med}(X)|$,

$$\hat{\theta} = \begin{cases} \text{MAD}/1.146 & \text{when } \hat{\alpha} = 2 \\ \text{MAD}/1.58 & \text{when } \hat{\alpha} = 3 . \end{cases}$$

Finally, the location estimator was obtained by a one step Newton-Raphson iteration of equation (14) starting with the median.

Comparison is made with the mean (normal case) and the sample median for the other three distributions. It is also compared with Tukey's bi-square, an M-estimator with $\eta(x) = x(1-x^2)^2$ for $|x| = 1$ and 0 otherwise. The value of $\lambda$ was taken to be $1/6.4\text{MAD}$, a value that is

observed by Bell (1980) and Johns to perform well over a wide range of distributions. For larger values of $\lambda$, the efficiency for the Cauchy and the slash improve, but the efficiency for the normal drops sharply.

Table 1 contains the results of a simulation. The small sacrifice in terms of the efficiency for the slash is exchanged for an improvement on the Laplace, the normal and the Cauchy. This is not meant to imply that estimation from this family should always be preferred to Tukey's bisquare. The strength of the bisquare rests in part on its treatment of situations not considered here. The implication intended is that model-based inference <u>can</u> be competitive even outside the family although it leaves open the question of whether the appropriate scheme is maximum likelihood.

TABLE I

SMALL SAMPLE EFFICIENCY OF ADAPTIVE ESTIMATES WITH RESPECT TO:

(A) THE MEAN OR MEDIAN

(B) TUKEY'S BISQUARE

Sample size: n = 20

Table gives estimated efficiency in %

UNDERLYING DISTRIBUTION

| | NORMAL | LAPLACE | CAUCHY | SLASH |
|---|---|---|---|---|
| (A) | 93# | 97* | 94* | 97* |
| (B) | 103.5 | 108 | 107 | 96.4 |
| Sample Size | 8000 | 20000 | 24000 | 24000 |

# efficiency relative to sample mean.

* efficiency relative to sample median.

Note: The "Princeton swindle" was used throughout. (cf. Andrews et al.). I thank Barry Eynon for running these simulations on the Stanford computer, and the Department of Statistics, Stanford University, for providing the computer time.

Modelling Dependence.

One possible use for this family of distributions is in modelling processes that are wider tailed than the normal and at the same time dependent. This may be needed, for example, in determining whether an estimate is robust against both correlation in the sample and mild departures from normality. Interactions between these two types of failures in the assumed model might be detected through Monte Carlo methods. There are two rather distinct ways in which a process having marginals (4) can fail to be independent. Unlike a Gaussian process, we may have $X_i$, $X_j$ uncorrelated and $X_i^2$, $X_j^2$ correlated for $i \neq j$. Modelling behaviour of this kind and testing estimates against this type of "second order" correlation may be important, especially since this is a type of dependence rarely checked for in practice. This kind of behaviour is by no means unusual. For example, consecutive changes in security prices (on a log scale) often seem to indicate no first order correlation but significant second order correlation.

We consider here processes analogous to the simpler Gaussian time series such as stationary first order autoregressive. For a process $\{X_t\}$ with marginals distributed as $Be(0,\theta,\alpha)$, we may define

$$X_{t+1} = B_t^{\frac{1}{2}} X_t + e_t \,. \tag{18}$$

The coefficients $B_t^{\frac{1}{2}}$ are not constant as in the Gaussian case, but are distributed as the square root of a beta variate with parameters $p\alpha$ and $(1-p)\alpha$ where $0 < p < 1$. The errors $e_t$ are distributed

as $Be(0,\theta,(1-p)\alpha)$ and the variables $B_t^{\frac{1}{2}}$, $X_t$, and $e_t$ are all independent. The correlation function of a process defined as in (18) is that of a stationary first order autoregressive:

$$\rho_h = \rho^{|h|} \quad \text{where} \quad \rho = EB_t^{\frac{1}{2}} = \frac{\Gamma(p\alpha+\frac{1}{2})\,\Gamma(\alpha)}{\Gamma(p\alpha)\,\Gamma(\alpha+\frac{1}{2})} . \tag{19}$$

Note that for a Gaussian process with correlation function (19), we would have $Cov(X_t^2,X_{t+h}^2) = \text{constant}\ \rho^{2|h|}$. In this case, however, we have a slower (exponential) rate of decay with the correlation between $X_t^2$ and $X_{t+h}^2$ given by $p^{|h|}$ (note that $p > \rho^2$).

A similar process can be defined with $B_t^{\frac{1}{2}}$ replaced by its negative, resulting in $\rho$ in (19) being replaced by its negative.

An alternative method of generating dependence (which may be used for the generation of a continuous time process) is the following: Let $Z_t$ be a stationary $(0,1)$ Gaussian sequence with correlation function $\psi(h)$. Let $G_t$ be a sequence of (possibly dependent) gamma $(\alpha,2\theta^2)$ variables, independent of the $Z_t$ sequence. Define:

$$X_t = G_t^{\frac{1}{2}} Z_t . \tag{20}$$

Then the autocovariance function of $X_t$ is:

$$Cov(X_t,X_{t+h}) = \psi(h)EG_t^{\frac{1}{2}} G_{t+h}^{\frac{1}{2}} \tag{21}$$

$$Cov(X_t^2,X_{t+h}^2) = (1+2\psi^2(h))E(G_tG_{t+h}) - 4\alpha^2\theta^4 . \tag{22}$$

Note that when $Z_t$ consists of i.i.d. normal variables, the first order correlation (21) is 0 but the second order (22) is not necessarily so. Therefore, this process may be used to model one such as the first difference in the logarithm of stock prices, which exhibit nearly normal behaviour, have no apparent correlation of the first order, but show a tendency for large values of $|X|$ to be followed by large values of $|X|$. All we need do to model such a process is introduce dependence in the $G_t$ sequence with either a moving average or an autoregressive type relation such as:

$$G_{t+1} = B_t G_t + \delta_t$$

where $B_t$, $G_t$, $\delta_t$ are independent variables having respectively the beta $(p\alpha, (1-p)\alpha)$, gamma $(\alpha, 2\theta^2)$ and gamma $((1-p)\alpha, 2\theta^2)$ distributions. This is the theme of [13].

A more convenient representation than (18) is available when $\alpha$ is an integer. In fact, a stationary sequence can be generated by the usual first order autoregressive relation:

$$X_{t+1} = \rho X_t + e_t \tag{23}$$

where $-1 < \rho < 1$ and $e_t$ is a random variable independent of $X_i$ having moment generating function:

$$m_e(s) = \left\{ \rho^2 + \frac{1-\rho^2}{1-\theta^2 s^2} \right\}^{\alpha} .$$

If we interpret $Be(0,\theta,\alpha)$ as a point mass at the origin whenever either $\theta$ or $\alpha = 0$, this is just the distribution of $Be(0,\theta,\alpha')$ where $\alpha'$ is distributed binomially with parameters $\alpha$ and $1-\rho^2$. Alternatively, it may be expressed as the sum of $\alpha$ random variables with distribution $Be(0,\theta',1)$ where $\theta' = 0$ or $\theta$ with probabilities $\rho^2$, $1-\rho^2$. Thus, this family of densities is in Feller's class L (Feller (1971), pp. 588-590).

Similarly, higher order autoregressive schemes are obtainable. For example, for a second order autoregressive process with distinct characteristic roots $\rho_1$ and $\rho_2$, both less than 1 in absolute value, the distribution of the errors is the distribution of the sum of $\alpha$ i.i.d. variates having distribution $Be(0,\theta',1)$ where $\theta'$ is a random variable assuming three values, $0$, $\rho_1\rho_2\theta$, and $\theta$.

Acknowledgement.

I am grateful to Barry Eynon for conducting the simulations, to D. Hinkley, V. Johns, and the referees for their help.

References


[1] Abramowitz, M., and Stegun, I. (1964). Handbook of Mathematical Functions, VIA Department of Commerce.

[2] Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., and Tukey, J. (1972). Robust Estimates of Location. Princeton University Press.

[3] Barndorff-Nielson, O. (1977). Exponentially Decreasing Distributions for the Logarithm of Particle Size. Proc. R. Soc. Lond. A. 353, 401-419.

[4] Bell, R. M. (1980). An Adaptive Choice of the Scale Parameter for M-estimators. Department of Statistics, Stanford University Technical Report No. 3.

[5] Bhattacharya, S. K. (1966). A Modified Bessel Function Model in Life Testing. Metrika 10, 133-144.

[6] Feller, W. (1971). An Introduction to Probability Theory and its Applications. Volume II, John Wiley and Sons, Inc., New York.

[7] Gradsheteyn, I. S., and Ryzhik, I. M. (1965). Tables of Integrals, Series, and Products. Academic Press.

[8] Holland, P. W., and Welsch, R. E. (1977). Robust Regression Using Iteratively Reweighted Least Squares. Comm. Statist. Theor. Meth. A6(9), 813-827.

[9] Johnson, N., and Kotz, S. (1970). Continuous Univariate Distributions 1 & 2. Houghton Mifflin.

[10] Laha, R. G. (1954). On Some Properties of the Bessel Function Distributions. Bull. Calc. Math. Soc. 46, 59-71.

[11] Lloyd, E. H., and Saleem, S. D. (1979). A Note on Seasonal Markov Chains with Gamma or Gamma-Like Distributions. J. Appl. Prob. 16, 117-128.

[12] McKay, A. T. (1932). A Bessel Function Distribution. Biometrika 24, 39-44.

[13] McLeish, D., and Pierson, H. A Dependent Increment Model for Security. Submitted.

[14] Pearson, K., Jeffery, G. B., and Elderton, E. M. (1929). On the Distribution of the First Product-Moment Coefficient in Samples Drawn from an Indefinitely Large Normal Population. Biometrika 21, 164-201.

[15] Pearson, K., Stouffer, S. A., and David, F. N. (1932). Further Applications in Statistics of the $T_m(x)$ Bessel Function. Biometrika 24, 316-343.

[16] Press, S. J. (1967). On the Sample Covariance from a Bivariate Normal Distribution. Ann. Inst. Stat. Math. 19, 355-361.

[17] Sichel, H. S. (1973). Statistical Valuation of Diamondiferous Deposit. J. S. Afr. Inst. Min. Metall. 73, 235-243.

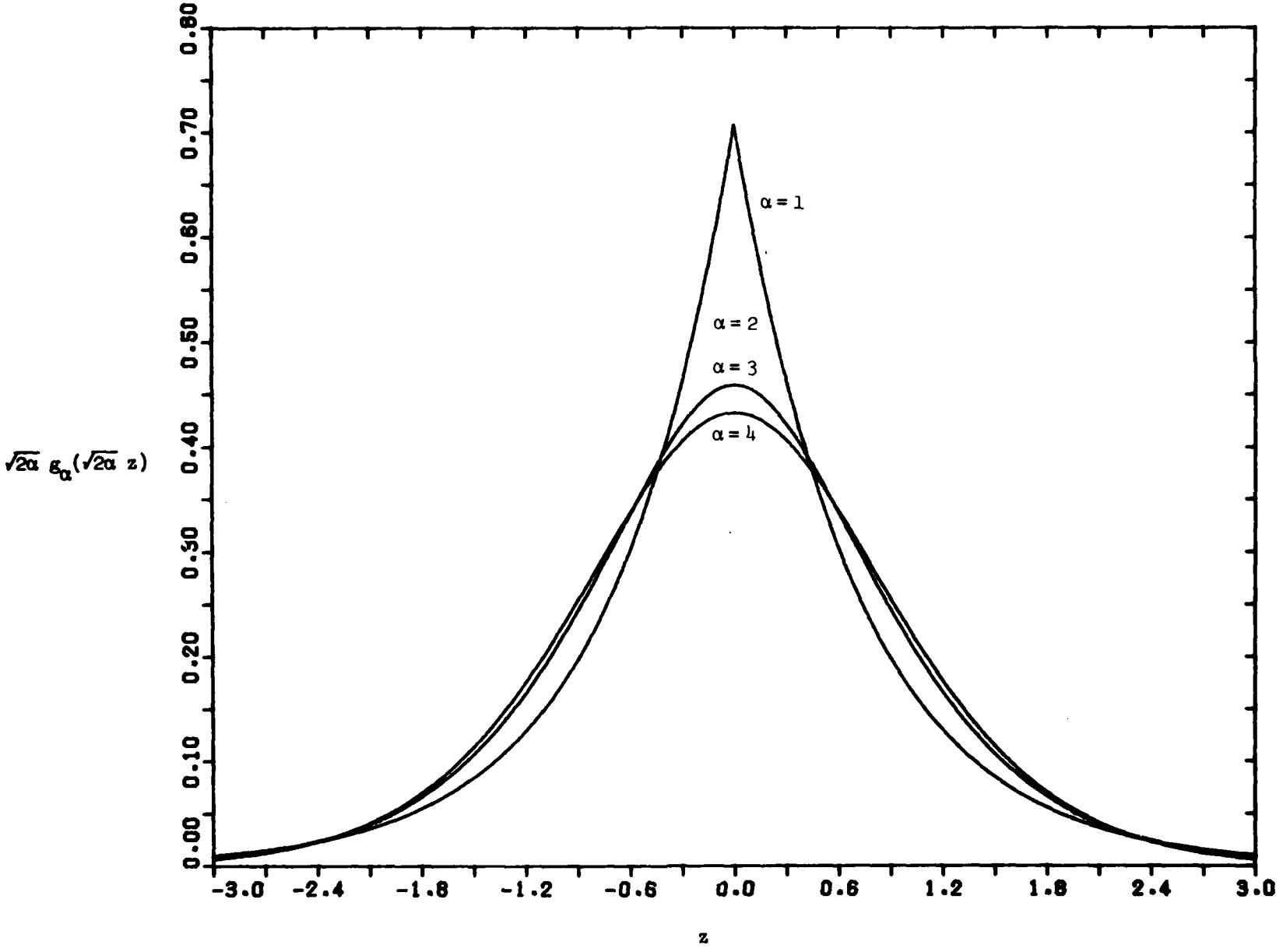
[18] Stigler, S. M. (1977). Do Robust Estimators Work with Real Data? Ann. Statist. 5, 1055-1098.

[19] Teichroew, D. (1957). The Mixture of Normal Distributions with Different Variances. Ann. Math. Statist. 28, 510-512.

[20] Wilkinson, G. N. (1979). Robust Inference - the Fisherian Approach. Robustness in Statistics, Academic Press, New York.

Figure 1

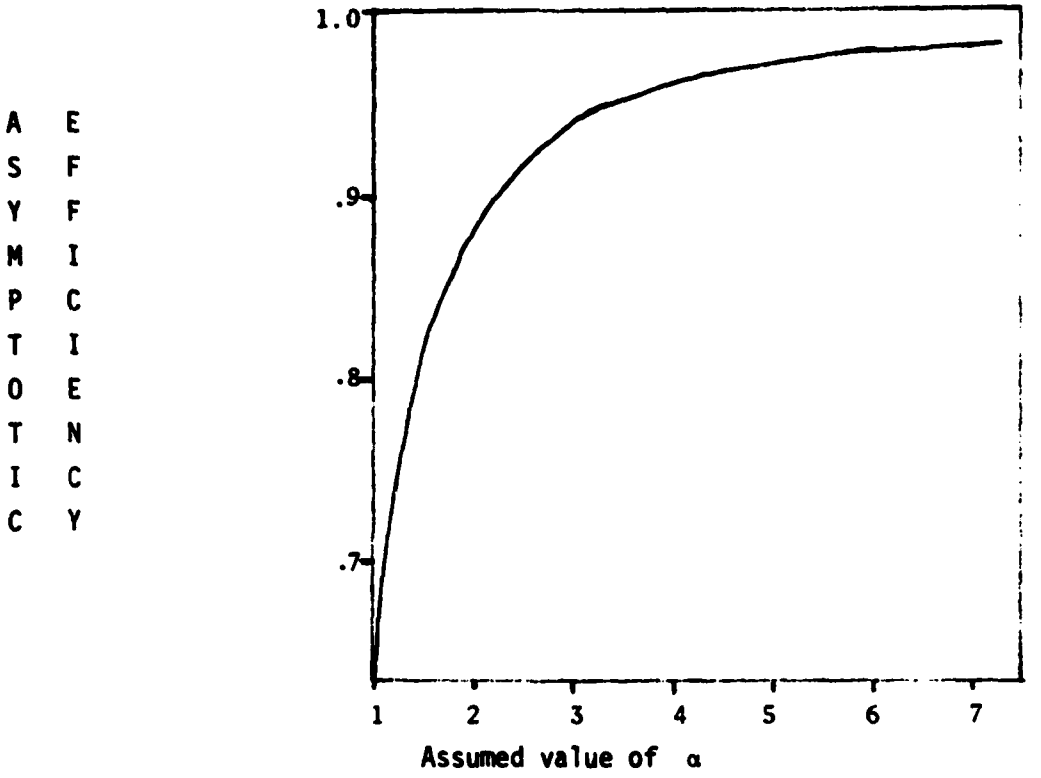
DENSITY OF BESSEL DISTRIBUTION




$\sqrt{2\alpha}\, g_\alpha(\sqrt{2\alpha}\, z)$

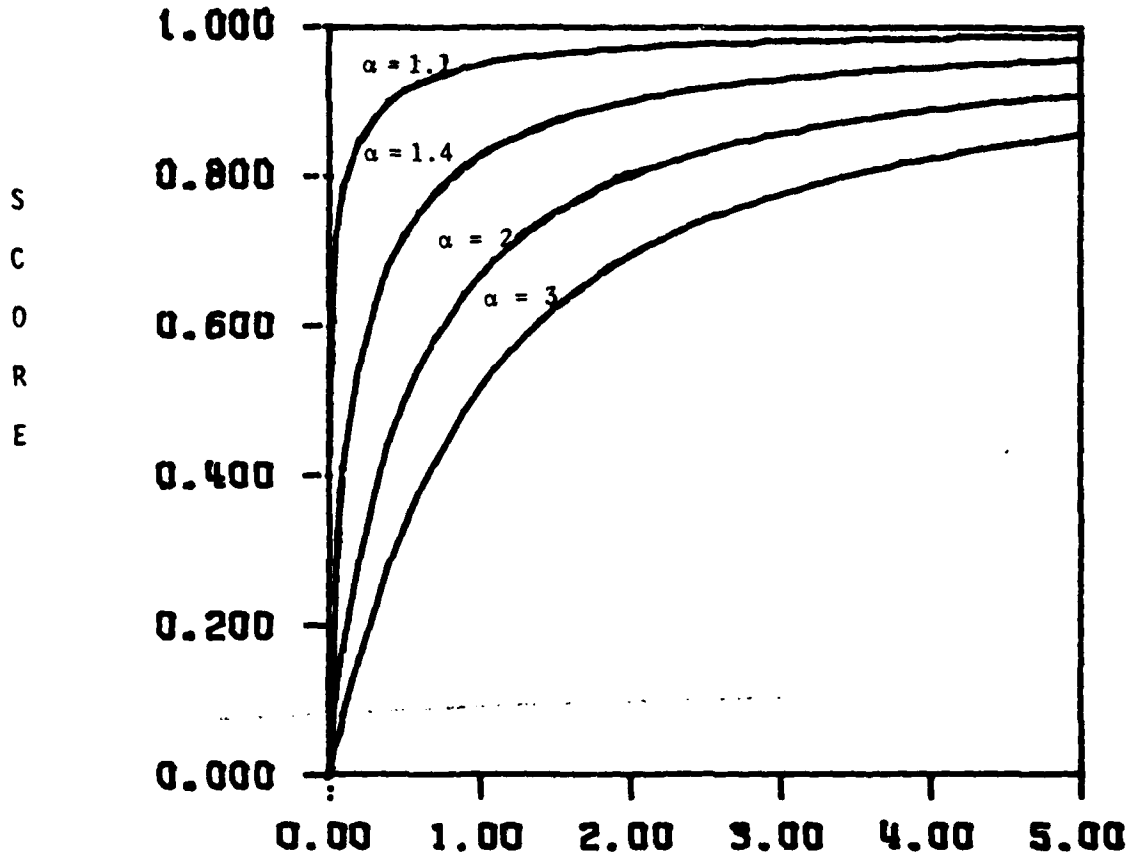
27

Figure 2




Asymptotic Efficiency of Estimate of Mean of Normal Sample for Assumed Value of $\alpha$

Figure 3




VALUE OF SCORE VS NUMBER OF STANDARD DEVIATIONS OF OBSERVATION FROM MEAN

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER 321 | 2. GOVT. ACCESSION NO. AD-A119378 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle) A ROBUST ALTERNATIVE TO THE NORMAL DISTRIBUTION | | 5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT |
| | | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s) D. L. McLeish | | 8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS Office Of Naval Research Statistics & Probability Program Code 411SP Arlington, VA 22217 | | 12. REPORT DATE July 7, 1982 |
| | | 13. NUMBER OF PAGES 29 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) | | 15. SECURITY CLASS. (of this report) UNCLASSIFIED |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Robust Estimation, Non-Normal Alternatives, Outliers.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE REVERSE SIDE.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

#321

A wider tailed family of distributions is suggested as an alternative to the normal distribution having many of the desirable properties of the normal family. One advantage of this alternative is the greater robustness of maximum likelihood estimates.